arXiv:2505.23709v1 [cs.CV] 29 May 2025

Skin Lesion Phenotyping via Nested Multi-modal Contrastive Learning

Dionysis Christopoulos[1], **Sotiris Spanos**[1], **Eirini Baltzi**[1],
Valsamis Ntouskos[2,1], **Konstantinos Karantzalos**[1]
[1] Remote Sensing Lab, National Technical University of Athens, Athens, Greece
[2] Department of Engineering and Sciences, Universitas Mercatorum, Rome, Italy

Abstract

We introduce SLIMP (Skin Lesion Image-Metadata Pre-training) for learning rich representations of skin lesions through a novel nested contrastive learning approach that captures complex relationships between images and metadata. Melanoma detection and skin lesion classification based solely on images, pose significant challenges due to large variations in imaging conditions (lighting, color, resolution, distance, etc.) and lack of clinical and phenotypical context. Clinicians typically follow a holistic approach for assessing the risk level of the patient and for deciding which lesions may be malignant and need to be excised, by considering the patient's medical history as well as the appearance of other lesions of the patient. Inspired by this, SLIMP combines the appearance and the metadata of individual skin lesions with patient-level metadata relating to their medical record and other clinically relevant information. By fully exploiting all available data modalities throughout the learning process, the proposed pre-training strategy improves performance compared to other pre-training strategies on downstream skin lesions classification tasks highlighting the learned representations quality.

1 Introduction

Categorizing skin lesions is an important part of dermatological examination, allowing clinicians to recognize potential skin malignancies and establish suitable follow-up actions and treatment plans. Among skin malignancies, melanoma, although having a lower incidence with respect to other skin cancers, such as basal cell carcinomas (BCCs), squamous cell carcinomas (SCCs) and other types of skin cancers, has a significantly heavier impact on the patient health in terms of morbidity and mortality. There are over 330,000 cases of melanoma diagnosed worldwide every year, leading to more than 55,000 deaths annually [1], with data suggesting an increased incidence in the last years [2]. Importantly, when detected early (stage I-II) melanoma can be cured in the majority of cases through surgical excision. This suggests the importance of developing efficient and effective methods for early detection of melanoma and other types of skin cancers.

Numerous works in the literature have attacked the problem of classifying skin lesions based on their appearance [3, 4], largely supported by the monumental effort put forward by the international skin imaging collaboration (ISIC) for constructing the ISIC datasets and organizing the corresponding challenges from 2016. In dermatological examination common practice, clinical decisions are not based solely on lesion appearance, but are also conditioned on additional characteristics of the lesions as well as patient phenotype and habits, as to alleviate overdiagnosis and wasteful use of resources. Drawing inspiration from this, recent datasets, including SLICE-3D [5], typically include lesion-only [6] or lesion and patient metadata [5, 7, 8].

Preprint.

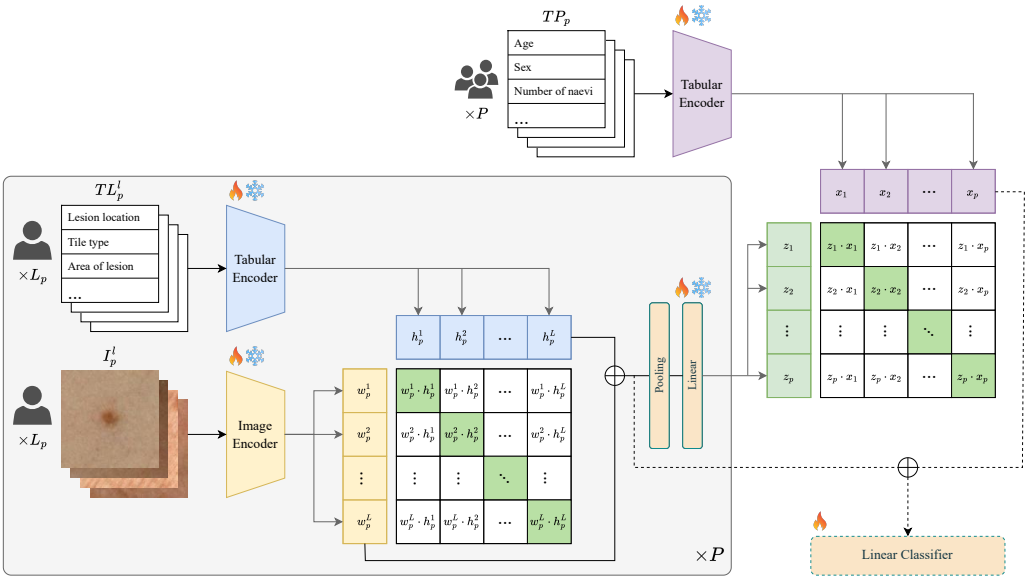

Figure 1: Architecture of the SLIMP approach. An inner multi-modal contrastive loss is employed to maximize agreement among images of skin lesions and the corresponding lesion-level metadata. Skin lesion image and metadata representations of a patient are aggregated, summarizing the lesion phenotype. At the patient level, agreement between the estimated lesion phenotype and their metadata is pursued through an outer contrastive loss.

Despite the significant effort dedicated in producing large collections of skin lesion data, still the data available are relatively scarce, due also to the difficulties in their collection and annotation making the development of deep-learning methods that rely on large data quantities troublesome. Suitable pre-text tasks offering self-supervision have proven to be invaluable in such scenarios, enabling the models to learn rich representative features that can be subsequently employed to address downstream tasks even when less data are available.

Building on these observations, we introduce SLIMP (Skin Lesion Image-Metadata Pre-training), a novel pre-training approach for skin lesions based on a nested multi-modal contrastive learning. SLIMP captures relations between the appearance of the lesions and the metadata associated with them in the context of the patient-level metadata. By incorporating both lesion and patient level metadata, the proposed method learns representative and generalizable features for skin lesions that can assist in downstream tasks. We specifically target to exploit all data modalities across all stages of the learning process. To enable effective transfer to target datasets with varying metadata, we employ an efficient continual pre-training approach for addressing the problems that arise from the differences that typically occur between the metadata available in different datasets. Additionally, by exploiting the similarity induced among the image and metadata features, we propose a method for enhancing datasets that do not contain metadata, by transferring metadata from a reference dataset with the target images based on the respective features, using their agreement in the shared embedding space.

The contributions of this work are the following: i) We propose a novel nested multi-modal pre-training strategy based on contrastive learning for producing rich skin lesion representations by leveraging metadata both at the lesion and patient levels, in relation to the lesion images; ii) We adapt the learned representations on different datasets through efficient continual pre-training, effectively addressing differences in metadata attributes, allowing to exploit metadata in all stages of the learning process; iii) We propose a retrieval strategy for enhancing image-only datasets using suitable reference metadata; iv) The proposed nested multi-modal pre-training strategy achieves improved skin lesion classification performance compared to reference pre-training strategies and strong baselines and competitive performance against supervised approaches.

2 Related work

Multi-modal self-supervised representation learning is used for enhancing image-based models by incorporating different data modalities, especially for tasks where additional context provides useful information for improved task performance.

CLIP [9], introduced a method for learning image-text representations through a contrastive learning paradigm. By linking each image to a natural language description, CLIP captures subtle patterns and nuances, creating representations that can accommodate different applications.

The work of Bourcier *et al.* [10] adopted a multi-modal pre-training approach for learning representations based on satellite imagery and associated metadata, showing that the additional context provided by metadata leads to improved performance in downstream tasks.

Regarding contrastive learning performed across taxonomies, [11] introduced hierarchical contrastive pre-training for images, allowing to consider labels organized in a taxonomy, by proposing a natural extension of the contrastive loss for hierarchical label relations as well as a constraint enforcing loss for separating distinct lineages. [12] used three levels of contrastive learning for improved sentiment analysis by incorporating various features combinations of the available data modalities.

In the medical domain, the work of [13] highlighted the importance of taking into account the patient-slide-patch hierarchy in learning suitable representations for cancer diagnosis based on whole-slide images. On the other hand, [14] used a contrastive loss spanning multiple levels across the same modality, ranging from patient-level to observation-level, for maximizing information utilization of the available data, leading to stronger representations for medical time-series analysis and classification.

In this work we adopt a contrastive learning strategy across two distinct levels of metadata, modeled as one level nested within the other, as patient-level metadata are shared while lesion-level metadata regard individual skin lesions. This scheme encourages learning of more representative skin-lesion representations that can assist in the downstream skin lesion classification task while offering improved generalization across different patients.

Continual pre-training has become a key strategy to make pretrained models more specialized and effective for real-world applications, where domain-specific knowledge is often crucial. In this context, [15] demonstrated that simply continuing to pretrain a language model on domain-specific texts substantially improves the accuracy across diverse tasks, even when labeled data is limited.

Lie *et al.* [16] developed a continual pre-training framework for the mBART model to boost machine translation for low-resource languages, where translation data is often limited or nonexistent. By generating mixed-language text from available monolingual resources, they enabled mBART to 'self-train' on noisy but representative data and extend its language skills to previously unseen languages.

In the domain of geospatial analysis, [17] tackled the resource-intense needs of geospatial applications with a continual pre-training method that exploits the rich representations coming from large-scale image datasets like ImageNet-22k.

The work of [18] extended this adaptive pre-training to general computer vision, aiming to address the high costs of self-supervised learning. Their approach, utilize existing pretrained models as a starting point to accelerate learning, achieving improved accuracy with fewer resources.

Multi-modal continual pre-training has only recently been explored, mainly regarding the adaptation of vision-language models [19, 20]. In the medical domain, [21] proposed continual pre-training for multi-modal medical data in a multi-stage manner to avoid interference between image and non-image modalities during learning.

The proposed method makes use of continual pre-training to fully exploit target dataset metadata. Due to the differences in the recorded attributes, continual pre-training allows adapting the metadata encoder accordingly, leading to improved classification performance. To the best of our knowledge, this is the first work that explores the use of multi-modal continual pre-training for tabular metadata, allowing to fully exploit the available metadata of target domains. Importantly, the proposed continual pre-training strategy does not rely on target labels, which are not always available in the context of skin lesion classification and other similar medical applications.

Data enhancement through retrieval has been proposed in the natural language processing domain under different settings. In [22], a retrieval-enhanced language model (RETRO) is introduced augmenting a frozen language model allowing retrieval from a large text database for improving its performance. In a similar direction, [23] proposed a discrete key-value bottleneck architecture considering pairs of sparse, separable and learnable key-value codes.

The work of [24] applies the idea in a multi-modal setting, establishing image-text correspondences using independently pre-trained image and text encoders by exploiting similarities within each modality in combination with a reduced dataset of known image-text correspondences. We consider a retrieval-enhanced variant of SLIMP for allowing multimodal classification even for image-only datasets, by matching metadata from a reference dataset.

3 Method

In this section we present SLIMP, a self-supervised pre-training approach with a nested contrastive loss. Given a reference skin-lesion classification dataset providing metadata at the lesion and at the patient levels, the proposed approach aims to learn representative and generalizable skin lesion representations by combining lesion images with the corresponding metadata at both levels. Two strategies are proposed for adapting these representations to target datasets in a way that fully exploit their metadata, even when the structure and content differ from the source data. This leads to enhanced performance on downstream tasks by leveraging multi-modal information about the skin lesions, as shown in Section 4. The notation used throughout this section is summarized in Table 6.

3.1 Nested constrastive multi-modal learning

The overall approach is presented in Figure 1 and summarized in Algorithm 1. For each patient $p \in \{1, ..., P\}$ our model process L_p lesion images $\{I_p^l\}_{l=1}^{L_p}$ with an image encoder to extract image-based features $\{w_p^l \in \mathbb{R}^D\}_{l=1}^{L_p}$, where D denotes the embedding size of each component. In parallel, the model processes the corresponding lesion-specific tabular metadata $\{TL_p^l\}_{l=1}^{L_p}$ with a tabular metadata encoder, to extract metadata-based feature representations $\{h_p^l \in \mathbb{R}^D\}_{l=1}^{L_p}$ on a lesion level. The resulting lesion-level representations are jointly optimized using an inner InfoNCE loss [25] in order to maximize their agreement. By maximizing the cosine similarity between the corresponding lesion image-metadata pairs and, analogously, minimizing the cosine similarity between non-matching pairs, the model learns a multi-modal lesion-level representations. The two lesion-level modalities are merged via concatenation, which has been shown to be a simple yet effective strategy [26] for obtaining a combined lesion-level representations $\{(w_p^l, h_p^l)\}_{l=1}^{L_p}$. These combined lesion representations are aggregated for all the lesions of a patient by applying average pooling and they are subsequently linearly transformed into a single vector $z_p \in \mathbb{R}^D$, summarizing the lesion phenotype of the patient. At the outer level, SLIMP processes the patient-specific tabular metadata (TP_p) utilizing an outer tabular metadata encoder, yielding a representation $x_p \in \mathbb{R}^D$. An outer InfoNCE loss is then applied between the patient-level metadata representation $x_p \in \mathbb{R}^D$ and the patient-level lesion phenotype representation $z_p \in \mathbb{R}^D$ obtained at the inner level. This nested contrastive pre-training framework enables the model to learn rich skin lesion representations that take into account the overall phenotype of the patient. Specifically, letting $s(\cdot, \cdot)$ denote the cosine similarity function and τ a temperature parameter, we employ a two-level nested contrastive loss with a weighting factor $\lambda \in [0, 1]$ as described below:

$$\mathcal{L}_{lesions}^{p} = -\frac{1}{2L_p} \sum_{l=1}^{L_p} \left(\log \frac{\exp(\mathrm{s}(w_p^l, h_p^l)/\tau)}{\sum_{j \in L_p} \exp(\mathrm{s}(w_p^l, h_p^j)/\tau)} + \log \frac{\exp(\mathrm{s}(h_p^l, w_p^l)/\tau)}{\sum_{j \in L_p} \exp(\mathrm{s}(h_p^j, w_p^l)/\tau)} \right), \quad (1)$$

$$\mathcal{L}_{patient} = -\frac{1}{2P} \sum_{p=1}^{P} \left(\log \frac{\exp(\mathrm{s}(z_p, x_p)/\tau)}{\sum_{i \in P} \exp(\mathrm{s}(z_p, x_i)/\tau)} + \log \frac{\exp(\mathrm{s}(x_p, z_p)/\tau)}{\sum_{i \in P} \exp(\mathrm{s}(x_i, z_p)/\tau)} \right), \quad (2)$$

$$\mathcal{L}_{total} = \frac{\lambda}{P} \sum_{p=1}^{P} \mathcal{L}_{lesions}^{p} + (1 - \lambda) \mathcal{L}_{patient}. \quad (3)$$

Algorithm 1: SLIMP pseudocode

Data: Lesion images: $\{\{I_p^l\}_{l=1}^{L_p}\}_{p=1}^{P}$, lesion metadata: $\{\{TL_p^l\}_{l=1}^{L_p}\}_{p=1}^{P}$, patient metadata: $\{TP_p\}_{p=1}^{P}$.

Sample a batch of B patients
$\mathcal{L}_{lesions} = 0$
for $p \in \{1, \ldots, B\}$ **do**
  Build batch of N lesion image-metadata pairs from patient p
  for $l \in \{1, \ldots, N\}$ **do**
    w_p^l = `ImageEncoder`(I_p^l)
    h_p^l = `LesionTabularEncoder`(TL_p^l)
  end
  $\mathcal{L}_{lesions} \mathrel{+}= \frac{1}{B}$`InfoNCELoss`$(\{w_p^l\}_{l=1}^{N}, \{h_p^l\}_{l=1}^{N})$
  z_p = `Linear(AvgPool(`$\{(w_p^l, h_p^l)\}_{l=1}^{N}$`))`
end
$\{x_p\}_{p=1}^{B}$ = `PatientTabularEncoder`$(\{TP_p\}_{p=1}^{B})$
$\mathcal{L}_{patient}$ = `InfoNCELoss`$(\{(z_p, x_p)\}_{p=1}^{B})$
$\mathcal{L}_{total} = \lambda \cdot \mathcal{L}_{lesions} + (1 - \lambda) \cdot \mathcal{L}_{patient}$

3.2 Handling divergent metadata of target datasets

SLIMP can be applied on reference, large-scale skin lesion classification datasets as [5] for learning lesion representation both from images and metadata. Nevertheless, due to differences in clinical practice, regulatory context, and other factors, metadata provided by different datasets, typically diverge in structure and/or collected attributes. To leverage all available data modalities on downstream tasks, we firstly propose an multi-modal continual pre-training approach for effectively adapting the learned representations to target datasets with potentially smaller size and diverging metadata. We also propose a retrieval-base strategy for allowing metadata-endowed skin lesion classification even for dataset which lack metadata completely.

Image-metadata continual pre-training When the target dataset provides metadata comprising different attributes and/or of different structure with respect to the reference one, a multi-modal continual pre-training approach on the target dataset is employed. In this case, the image and tabular encoders are fine-tuned to adapt the representations on the input features of the target domain. This requires for the target dataset to contain both lesion-level and patient-level metadata. In cases where patient metadata are unavailable, a variant of this setup is considered, using the lesion level loss alone, taking into account solely the lesion images and the corresponding metadata. This allows to cope with varying levels of data availability while maintaining robust continual feature learning across both the image and metadata modalities.

Dataset enhancement via metadata retrieval When the target dataset lacks metadata, a retrieval-based approach is followed for artificially enhancing the target dataset by creating metadata pseudo-modalities. As lesion metadata are tightly related to the corresponding images, we consider the possibility of enhancing datasets which do not provide metadata by constructing pseudo-modalities of patient-level and lesion-level metadata using the corresponding modalities of the reference dataset on which the SLIMP model has been pre-trained. Drawing inspiration from [24], and building on the fact that the lesion and patient level modalities have been trained to maximize agreement, we use the encoding of the lesion images to retrieve the metadata of the original dataset that exhibit the highest similarity and use them on downstream tasks 'as-if' they were accompanying metadata.

Specifically, in this setup, presented in Figure 2, the model utilizes only the images I_p^l from the target dataset, passing them through the image encoder of the SLIMP model that has been pre-trained on the reference dataset, providing the target dataset image representations w_p^l. Based on these features, we then conduct a two-step retrieval process to incorporate additional context from the reference dataset metadata representations. First, we compare w_p^l with the features $\tilde{h}^l$ derived from the pre-trained SLIMP lesion metadata encoder and we retrieve the vector $\hat{h}^l$ with the highest cosine similarity. The combined feature set $\{(w_p^l, \hat{h}^l)\}$ is linearly transformed into a single patient-level vector $\hat{z}_p$, which is then compared with the features $\tilde{x}_p$ derived from the pre-trained SLIMP patient metadata

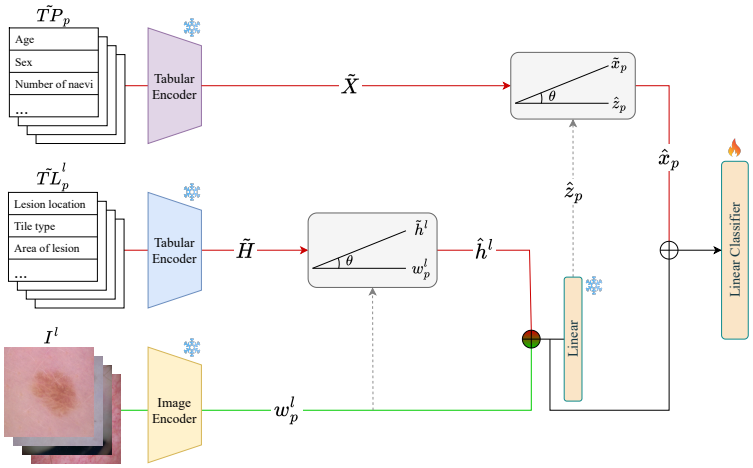

Figure 2: Retrieval of lesion and patient metadata from the reference dataset (red path) for constructing metadata pseudo-modalities for the target dataset (green path) using SLIMP. (Best viewed zoomed-in)

encoder, to retrieve the most relevant $\hat{x}_p$. By adding pseudo-modalities on both the patient and the lesion-level, this retrieval process produces three feature vectors for each image of the target dataset $\hat{y}_p^l : \{(w_p^l, \hat{h}^l, \hat{x}_p)\}$ that can be used for lesion classificition in the target datasets.

4 Experimental evaluation

4.1 Datasets

Evaluation is performed considering five widely used, public skin lesion datasets, which differ in key aspects, including dataset size, image type (dermoscopic or clinical), availability of metadata (such as the number of patient clinical features), and degree of class imbalance. Namely, the datasets considered are: SLICE-3D[5], PAD-UFES-20 [7], HIBA [27], HAM10000 [8], and PH2 [6]. Their main characteristics are summarized in Table 1. The skin lesion classification datasets contain different taxonomies, with important class imbalance of varying degrees (Figure 3). To allow comparison across all the datasets, we mainly consider the task of classifying lesions in benign and malignant. Section B provides additional dataset details and benign and malignant class definition.

4.2 Implementation

Unless otherwise stated, we employ ViT-Small [28] as a transformer-based image encoder and TRACE [29] as a transformer-based tabular data encoder. We train the model for 150 epochs on an NVIDIA RTX A6000 GPU with 48GB of VRAM. For pre-training the model on the SLICE-3D dataset, we consider a batch size of 4 patients and $N = 100$ lesions. For continual pre-training on target datasets, we fine-tune the embedding layers of the image encode and learn new metadata encoders for the metadata encoder, keeping the attention layers of all encoders frozen. We have observed that this strategy leads to increased downstream performance. During continual pre-training, the batch size is increased to 64 patients. For allowing downstream performance assessment we randomly split

Dataset	Number of Samples	Number of Patients	Targets	Patient Missing Values(%)	Lesion Missing Values(%)	Patient Metadata	Lesion Metadata
SLICE-3D	401,059	1,042	Benign/Malignant	1.78%	0.04%	✓	✓
PAD-UFES-20	2,298	1,373	Multiclass	32.2%	7.00%	✓	✓
HIBA	1,616	623	Multiclass	21.2%	12.8%	✓	✓
HAM10000	10,015	N/A	Multiclass	0.57%	1.17%	✓	✓
PH2	200	N/A	Multiclass	N/A	6.12%	✗	✓

Table 1: Main aspects of skin lesion datasets considered in the evaluation.

the target datasets into training and validation splits with a ratio of 90%-10%, respectively. Both pre-training stages use the Adam optimizer with a learning rate of 10^{-4} and $\lambda = 0.9$.

4.3 Protocol

The SLIMP model is pre-trained on SLICE-3D, a large-scale medical imaging dataset. As per standard practice [30], for evaluating the intrinsic quality of the learned features learned by the SLIMP model through self-supervision, evaluation is performed by considering k-nearest neighbors (kNN) with $k = 10$ and linear probing classification on the downstream skin-lesion classification task on different target datasets. The impact of SLIMP features stemming from different modalities is also analyzed in the results. Performance of the models is evaluated considering four metrics: Accuracy (Acc), Balanced Accuracy (BAcc), F1-Score, and area under receiver operator curve (AUC). Balanced Accuracy corresponds to the average of the Sensitivity and Specificity scores and is particularly relevant in the medical domain as it captures the model's ability to correctly identify positive and negative instances, even when datasets suffer from significant class imbalance.

4.4 Results

Our main goal is to assess the quality of the skin lesion representations learned by the proposed SLIMP model. Additionally, we examine the extent in which the use of metadata in different parts of the pipeline impacts the performance on the downstream task. In these regards, we consider strong baselines in each of these parts. Table 2 summarizes the results.

We first consider comparison using features that have been obtained via pre-training on the reference SLICE3D dataset. In this context, we consider the Pre-SLIMP setup, which uses the visual features extracted by the image encoder of SLIMP pre-trained on the lesion and patient metadata of SLICE3D, and compare it against the features obtained by SimCLR [31] pre-trained on the images of SLICE3D. We also consider the downstream classification performance of the subclass-balancing contrastive learning approach (SBCL) proposed in [32]. We observe that Pre-SLIMP shows competitive performance with SimCLR, even though it does not consider any image self-supervision, surpassing it in PAD-UFES-20 and HAM10000, while showing lower performance in HIBA and PH2. This suggests that SLIMP incorporates information from corresponding metadata in the image representation, potentially enabling to be robust against image domain shift. Similar conclusions can be drawn from the kNN classification results presented in Table 3.

Pre-SLIMP also outperforms SBCL, which explicitly handles class imbalance and long-tail distributions. Additionally, Table 2 provides the results from the MAE [33], BeiTv2 [34] and DINOv2 [35] generic foundation models. For a qualitative evaluation, attention maps produced by the SLIMP method are compared against the ones obtained by DINOv2 and MAE, and SimCLR in Section F.

Use of metadata We note that, as the metadata attributes of the target datasets differ from the reference one, the pre-trained metadata encoders cannot be directly used. This shortcoming is addressed by the SLIMP model, which applies continual pre-training on the target dataset as described in Section 4.2. This allows the use of target dataset metadata, both at the continual pretraining stage and at the downstream classification task. We see that the image representations obtained after continual pre-training, denoted as $\text{SLIMP}_{\text{IMAGE}}$, improve downstream classification performance, clearly outperforming the SBCL method continually pre-trained on the target datasets (SBCL-C). Importantly, the complete SLIMP method, which uses the features obtained by all data modalities in the downstream task, leads to improved performance across all the metrics for most datasets, both for linear probing (Table 2) and kNN (Table 3) evaluation. Interestingly, SLIMP also shows competitive performance compared to TFormer [36], a fully supervised model for multi-modal lesion classification trained directly on both the images and metadata of the target dataset, showing a decrease in performance only for the PAD-UFES-20 dataset.

To assess the effectiveness of the nested-architecture in providing better skin lesion representations, we also consider a variant of SLIMP, $\text{SLIMP}_{\text{FLAT}}$, which comprises a single InfoNCE loss, applied between the image features and the features obtained by a tabular encoder operating on the concatenated patient-lesion metadata. SLIMP clearly outperforms this single-level variant, demonstrating the effectiveness of its nested contrastive learning architecture in capturing image-metadata relations.

		PAD-UFES-20				HIBA				HAM10000				PH2			
	MD	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC
Generic Pretrained Models																	
MAE	✗	51.2	51.7	0.498	0.486	45.1	47.7	0.582	0.289	82.7	64.7	0.441	0.863	-	-	-	-
BEiTv2	✗	56.5	54.1	0.462	0.491	54.3	50.6	0.362	0.412	80.5	50.0	0.445	0.516	80.0	50.0	0.444	0.914
DINOv2	✗	75.2	75.0	0.750	0.783	75.3	74.3	0.744	0.799	84.2	71.0	0.727	0.879	86.7	81.3	0.785	0.867
Pretrained on SLICE3D																	
SimCLR	✗	70.4	70.5	0.699	0.766	84.6	84.3	0.863	0.913	81.2	69.4	0.522	0.868	**95.0**	87.5	0.857	**1.000**
SBCL	✗	66.1	66.0	0.642	0.672	66.7	67.5	0.671	0.671	56.0	63.8	0.403	0.710	75.0	75.0	0.546	0.734
Pre-SLIMP	✗	78.3	78.2	0.793	0.816	75.3	75.5	0.747	0.863	82.9	66.3	0.474	0.851	90.0	83.3	0.800	0.941
Ret-SLIMP	✓*	78.3	78.4	0.795	0.810	77.8	78.0	0.772	0.851	84.3	69.7	0.534	0.857	90.0	88.1	0.833	0.952
Continual pre-training																	
SBCL-C	✗	71.3	71.1	0.689	0.711	72.2	73.9	0.762	0.760	62.2	73.4	0.484	0.816	90.0	84.4	0.750	0.719
$\text{SLIMP}_{\text{IMAGE}}$	✗	76.1	75.5	0.764	0.807	77.8	78.1	0.763	0.867	84.7	69.2	0.529	0.889	**95.0**	**96.4**	**0.923**	0.988
$\text{SLIMP}_{\text{FLAT}}$	✓	85.7	85.3	0.872	0.906	84.6	84.5	0.854	0.911	84.4	75.6	0.608	0.894	**95.0**	**96.4**	**0.923**	0.988
SLIMP	✓	90.9	90.2	**0.921**	0.926	**90.7**	**90.6**	**0.914**	0.947	**85.9**	**78.5**	**0.650**	**0.901**	**95.0**	**96.4**	**0.923**	0.988
Supervised																	
TFormer	✓	**91.3**	**91.3**	0.917	**0.960**	88.9	88.9	0.892	**0.963**	82.1	76.2	0.601	0.875	**95.0**	91.7	0.909	0.988
Low-shot Evaluation																	
SLIMP 1%	✓	83.9	84.1	0.849	0.908	75.3	75.8	0.726	0.863	76.0	69.6	0.493	0.802	70.0	64.3	0.500	0.548
SLIMP 5%	✓	84.4	84.1	0.858	0.912	80.9	81.1	0.803	0.901	76.8	71.0	0.512	0.821	90.0	83.3	0.800	0.952
SLIMP 10%	✓	88.7	88.2	0.901	0.922	84.0	84.2	0.835	0.917	77.0	72.3	0.526	0.818	90.0	84.5	0.833	0.952

Table 2: Comparison of SLIMP with various baselines, on the lesion classification task using linear probing. MD stands for 'Metadata' used for downstream classification, while asterisk (*) denotes the use of metadata from the reference dataset (SLICE3D). For all metrics higher values are better. Best results are in **bold**, second best are underlined.

		PAD-UFES-20				HIBA				HAM10000				PH2			
	MD	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC
Pretrained on SLICE3D																	
SimCLR	✗	67.4	67.1	0.631	0.765	80.3	80.1	0.822	0.902	**87.2**	69.7	0.552	0.935	85.0	62.5	0.400	0.914
Pre-SLIMP	✓	70.0	70.4	0.706	0.752	72.8	72.9	0.732	0.847	82.0	62.5	0.400	0.818	90.0	83.3	0.800	0.941
Ret-SLIMP	✓*	72.2	72.3	0.738	0.769	72.8	73.2	0.707	0.815	80.8	59.9	0.343	0.789	95.0	91.7	0.909	0.887
Continual																	
$\text{SLIMP}_{\text{IMAGE}}$	✓	70.9	70.4	0.739	0.764	74.1	74.5	0.716	0.818	82.6	64.4	0.439	0.861	95.0	96.4	0.923	0.988
$\text{SLIMP}_{\text{FLAT}}$	✓	81.3	81.5	0.823	0.881	77.8	77.8	0.783	0.895	84.1	73.0	0.576	0.884	95.0	96.4	0.923	0.988
SLIMP	✓	**89.6**	**89.6**	**0.904**	**0.927**	**87.7**	**87.5**	**0.886**	**0.924**	85.9	**75.2**	**0.618**	**0.888**	95.0	96.4	0.923	0.988

Table 3: Comparison of SLIMP with baselines on the lesion classification task using kNN with $k = 10$. MD stands for 'Metadata' used for downstream classification, while asterisk (*) denotes the use of metadata from the reference dataset (SLICE3D). For all metrics, higher values are better. Best results are in **bold**, second best are underlined.

The use of pseudo-modalities constructed through retrieval of metadata from the reference dataset, denoted as Ret-SLIMP in the tables, shows consistently improved performance compared to Pre-SLIMP and comparable performance with $\text{SLIMP}_{\text{IMAGE}}$, even though it has not seen any data from the target datasets during training. This is valuable when the target dataset lacks metadata. This observation also further highlights the importance of using metadata for downstream classification.

Low-shot evaluation The proposed multi-modal continual pre-training strategy does not rely on target labels. This is crucial as data labeling is expensive and time-consuming, especially in the context of skin lesion classification and other similar medical applications. To further assess the quality of the learning representations, we examine how SLIMP performs in a low-shot learning setting, considering 1 %, 5 %, and 10 % of the target dataset labels for downstream classification. The results, presented in the last rows of Table 2 (highlighted in orange), indicate that the SLIMP features lead to remarkable low-shot learning performance. It is interesting to note that in most cases, SLIMP low-shot performance is better than $\text{SLIMP}_{\text{IMAGE}}$ and $\text{SLIMP}_{\text{FLAT}}$. The first suggests the importance of the model making use of metadata both during pre-training, but also for the downstream

Method	Metadata	Acc	F1-macro	F1-weighted
SBCL	✗	45.7	0.289	0.433
SimCLR	✗	**84.2**	0.688	0.826
TFormer	✓	78.7	0.698	0.792
SLIMP	✓	82.9	**0.825**	**0.835**

Table 4: Comparison of SLIMP with SBCL, SimCLR and TFormer baselines for an imbalanced multiclass classification task on PAD-UFES-20 dataset. Best results in **bold**.

Image	Metadata		PAD-UFES-20				HIBA				HAM10000				PH2			
	Lesion	Patient	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC
			Linear Probing															
✓	✗	✗	76.1	75.5	0.764	0.807	77.8	78.1	0.763	0.867	84.7	69.2	0.529	0.889	-	-	-	-
✓	✓	✗	88.3	88.2	0.892	0.917	87.0	86.7	0.885	0.916	84.0	69.3	0.527	0.882	95.0	96.4	0.923	0.988
✓	✓	✓	**90.9**	**90.2**	**0.921**	**0.926**	**90.7**	**90.6**	**0.914**	**0.947**	**85.9**	**78.5**	**0.650**	**0.901**	-	-	-	-
			kNN															
✓	✗	✗	70.9	70.4	0.739	0.764	74.1	74.5	0.716	0.818	82.6	64.4	0.439	0.861	-	-	-	-
✓	✓	✗	87.0	87.2	0.877	0.919	81.5	81.4	0.824	0.905	85.1	70.8	0.555	**0.888**	95.0	96.4	0.923	0.988
✓	✓	✓	**89.6**	**89.6**	**0.904**	**0.927**	**87.7**	**87.5**	**0.886**	**0.924**	**85.9**	**75.2**	**0.618**	**0.888**	-	-	-	-

Table 5: Ablation study of the SLIMP encoder outputs used for downstream classification.

classification task. Comparable performance to $\text{SLIMP}_{\text{FLAT}}$ further highlights the ability of the nested contrastive learning to capture relations among metadata and images.

Multiclass classification In Table 4 we evaluate our proposed SLIMP method in a multiclass classification setting on PAD-UFES-20 dataset, in comparison with SimCLR, SBCL, and TFormer. We report results for the overall Accuracy (Acc), F1-macro (which ensures equal contribution from minority classes), and F1-weighted (which accounts for class imbalance). Notably, SLIMP outperforms all baselines in both F1-score metrics, highlighting the robustness of SLIMP in handling imbalanced multiclass classification tasks. We note that techniques addressing class imbalance can be combined with SLIMP to further improve multiclass classification performance.

4.5 Ablation

To assess the importance of incorporating two distinct levels of metadata, we compare different variants of SLIMP in Table 5. Specifically, in the first row of we consider the linear probing performance of a variant where only the image encoder is fine-tuned on the target dataset, while in the second row we consider a variant where the image encoder is still fine-tuned on the target dataset, while a single metadata encoder is used, trained on the lesion metadata of the target dataset alone. The third row shows the results of the proposed SLIMP model. The last three rows report analogous results with kNN classification. The results suggest that the addition of each modality contributes positively to the downstream task performance. Additional ablations are provided in Section C.

5 Conclusions and limitations

We have presented SLIMP, a novel nested multi-modal pre-training strategy for learning rich skin lesion representations by considering lesion images in combination with associated lesion-level as well as patient-level metadata. The experimental evaluation demonstrates SLIMP's ability to learn representations that improve performance in downstream classification tasks, by combining information about the patient's lesion phenotype, with information regarding their traits and habits. In this context, we propose strategies for fully exploiting available metadata, through all the stages of the learning process, including a method that enables the enhancement of image-only skin lesion datasets by 'borrowing' patient and lesion metadata from reference pre-training data. Importantly, the proposed method does not rely on data annotations, handling a major challenge in healthcare applications where data annotation incurs significant costs. The results obtained for low-shot settings of the target datasets, demonstrate the quality of the obtained skin lesion representations as they enable high classification performance even with minimal labeled data. Considering the above, our proposed method has the potential to become widely applicable in clinical settings, providing insights and decision support during skin lesion diagnosis.

Despite its strengths, the proposed method has certain limitations. Firstly, the nested pre-training strategy requires a data structure that incorporates both patient- and lesion-level metadata, which may limit its adaptability to other domains where such structured scenarios do not straight-forwardly exist. Secondly, significant shift in the image domain, including high variability in the sources and resolutions of lesion images, can possibly downgrade downstream performance. This problem can be addressed by incorporating image augmentations in the learning process. Regarding negative impacts, it should be noted that misuse of this method, as for all computer-aided diagnosis methods, can lead to overdiagnoses, or misdiagnoses, with important psychological and economic repercussions. Hence, real-life use of such systems should be intended only for assisting the decision-making of expert users, and not for direct use by the patients.

Acknowledgments and Disclosure of Funding

This work was supported by the iToBoS EU H2020 project under Grant 965221. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

References

[1] M. Arnold, D. Singh, M. Laversanne, J. Vignat, S. Vaccarella, F. Meheus, A. E. Cust, E. de Vries, D. C. Whiteman, and F. Bray, "Global burden of cutaneous melanoma in 2020 and projections to 2040," *JAMA Dermatology*, vol. 158, no. 5, pp. 495–503, 05 2022.

[2] Y. Sun, Y. Shen, Q. Liu, H. Zhang, L. Jia, Y. Chai, H. Jiang, M. Wu, and Y. Li, "Global trends in melanoma burden: A comprehensive analysis from the global burden of disease study, 1990-2021," *Journal of the American Academy of Dermatology*, 2024.

[3] M. K. Hasan, M. A. Ahamad, C. H. Yap, and G. Yang, "A survey, review, and future trends of skin lesion segmentation and classification," *Computers in Biology and Medicine*, vol. 155, p. 106624, 2023.

[4] A. Adegun and S. Viriri, "Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art," *Artificial Intelligence Review*, vol. 54, no. 2, pp. 811–841, 2021.

[5] N. Kurtansky, B. D'Alessandro, M. Gillis, B. Betz-Stablein, S. Cerminara, R. Garcia, E. Goessinger, P. Gottfrois, P. Guitera, A. Halpern, V. Jakrot, H. Kittler, K. Kose, K. Liopyris, J. Malvehy, V. Mar, L. Martin, T. Mathew, and V. Rotemberg, "The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection," *Scientific Data*, vol. 11, 08 2024.

[6] T. Mendonça, P. Ferreira, J. Marques, A. Marçal, and J. Rozeira, "PH2 - A dermoscopic image database for research and benchmarking," in *IEEE Engineering in Medicine and Biology Society*, 2013, pp. 5437–5440.

[7] A. G. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. Alves Jr, J. G. Esgario, A. C. Simora, P. B. Castro, F. B. Rodrigues, P. H. Frasson, R. A. Krohling, H. Knidel, M. C. Santos, R. B. do Espírito Santo, T. L. Macedo, T. R. Canuto, and L. F. de Barros, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, p. 106221, 2020.

[8] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, 2018.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[10] J. Bourcier, G. Dashyan, K. Alahari, and J. Chanussot, "Learning representations of satellite images from metadata supervision," in *European Conference on Computer Vision*, 2024, pp. 1–30.

[11] S. Zhang, R. Xu, C. Xiong, and C. Ramaiah, "Use all the labels: A hierarchical multi-label contrastive learning framework," in *Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 660–16 669.

[12] C. Fan, K. Zhu, J. Tao, G. Yi, J. Xue, and Z. Lv, "Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, pp. 1–17, 2024.

[13] C. Jiang, X. Hou, A. Kondepudi, A. Chowdury, C. W. Freudiger, D. A. Orringer, H. Lee, and T. C. Hollon, "Hierarchical discriminative learning improves visual representations of biomedical microscopy," in *Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 798–19 808.

[14] Y. Wang, Y. Han, H. Wang, and X. Zhang, "Contrast everything: A hierarchical contrastive framework for medical time-series," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, 2023, pp. 55 694–55 717.

[15] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 8342–8360.

[16] Z. Liu, G. I. Winata, and P. Fung, "Continual mixed-language pre-training for extremely low-resource neural machine translation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2706–2718.

[17] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *International Conference on Computer Vision*, October 2023, pp. 16 806–16 816.

[18] C. J. Reed, X. Yue, A. Nrusimha, S. Ebrahimi, V. Vijaykumar, R. Mao, B. Li, S. Zhang, D. Guillory, S. Metzger, K. Keutzer, and T. Darrell, "Self-supervised pretraining improves self-supervised pretraining," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, January 2022, pp. 2584–2594.

[19] K. Roth, V. Udandarao, S. Dziadzio, A. Prabhu, M. Cherti, O. Vinyals, O. Hénaff, S. Albanie, M. Bethge, and Z. Akata, "A practitioner's guide to continual multimodal pretraining," *arXiv preprint arXiv:2408.14471*, 2024.

[20] Y. Chen, L. Meng, W. Peng, Z. Wu, and Y.-G. Jiang, "Comp: Continual multimodal pre-training for vision foundation models," *arXiv preprint arXiv:2503.18931*, 2025.

[21] Y. Ye, Y. Xie, J. Zhang, Z. Chen, Q. Wu, and Y. Xia, "Continual self-supervised learning: Towards universal multi-modal medical data representation learning," in *Computer Vision and Pattern Recognition*, 2024, pp. 11 114–11 124.

[22] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark *et al.*, "Improving language models by retrieving from trillions of tokens," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2206–2240.

[23] F. Träuble, A. Goyal, N. Rahaman, M. Mozer, K. Kawaguchi, Y. Bengio, and B. Schölkopf, "Discrete key-value bottleneck," in *International Conference on Machine Learning*, 2023, pp. 34 431–34 455.

[24] A. Norelli, M. Fumero, V. Maiorca, L. Moschella, E. Rodolà, and F. Locatello, "Asif: Coupled data turns unimodal models to multimodal without training," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 15 303–15 319.

[25] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] W.-H. Weng, Y. Cai, A. Lin, F. Tan, and P.-H. C. Chen, "Multimodal multitask representation learning for pathology biobank metadata prediction," *arXiv preprint arXiv:1909.07846*, 2019.

[27] International Skin Imaging Collaboration (ISIC), "ISIC Archive Collection 176," 2024. [Online]. Available: https://api.isic-archive.com/collections/176/

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[29] D. Christopoulos, S. Spanos, V. Ntouskos, and K. Karantzalos, "TRACE: Transformer-based Risk Assessment for Clinical Evaluation," *arXiv preprint arXiv:2411.08701*, 2024.

[30] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *International Conference on Computer Vision (ICCV)*, 2021, pp. 9650–9660.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607.

[32] C. Hou, J. Zhang, H. Wang, and T. Zhou, "Subclass-balancing contrastive learning for long-tailed recognition," in *ICCV*. IEEE, 2023, pp. 5372–5384.

[33] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 15 979–15 988. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01553

[34] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "Beit v2: Masked image modeling with vector-quantized visual tokenizers," *CoRR*, vol. abs/2208.06366, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.06366

[35] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[36] Y. Zhang, F. Xie, and J. Chen, "Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis," *Computers in Biology and Medicine*, vol. 157, p. 106712, 2023.

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 10 347–10 357.

[38] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 18 932–18 943.

[39] C. Patrício, L. F. Teixeira, and J. C. Neves, "Towards concept-based interpretability of skin lesion diagnosis using vision-language models," in *ISBI*. IEEE, 2024, pp. 1–5.

[40] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2020.

[41] K. Cao, C. Wei, A. Gaidon, N. Aréchiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019, pp. 1565–1576.

[42] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, S. Becker, S. Thrun, and K. Obermayer, Eds., vol. 15. MIT Press, 2002.

A Notation

Table 6 summarizes the notation used throughout the manuscript.

Notation	Description
P	Number of patients indexed by $p \in \{1, ...P\}$
L_p	Total lesions of patient p indexed by $l \in \{1, ...L_p\}$
TP_p	Tabular metadata for patient p
TL_p^l	Tabular metadata for lesion l of patient p
I_p^l	Lesion image l of patient p
w_p^l	Image encoder output of I_p^l
h_p^l	Tabular encoder output of TL_p^l
x_p	Tabular encoder output of TP_p
z_p	Linearly transformed output based on $\{w_p^l, h_p^l\}$
D	Output dimensionality of each component
$\tilde{H} = \{\tilde{h}^l\}_{l=1}^L$	Lesion-level pre-trained features of original dataset
$\tilde{X} = \{\tilde{x}_p\}_{p=1}^P$	Patient-level pre-trained features of original dataset
$\hat{h}^l$	Retrieved features from $\tilde{H}$
$\hat{z}_p$	Linearly transformed output based on $\{w_p^l, \hat{h}_p^l\}$
$\hat{x}_p$	Retrieved features from $\tilde{X}$
$\hat{y}_p^l$	$\text{concat}\{w_p^l, \hat{h}^l, \hat{x}_p\}$

Table 6: Summary of the notation.

B Dataset details

The following skin-lesion classification datasets are considered:

SLICE-3D [5]: a public skin lesion dataset containing up to 401,059 15mm-by-15mm field-of-view cropped images, centered on distinct lesions extracted from 3D Total Body Photography (TBP) collected across seven dermatologic centers worldwide. The dataset was curated for the ISIC 2024 Challenge and contains 40 clinical features concerning both patients and lesions, such as age, sex, general anatomic site, common patient identifier, clinical size, and various data fields from the TBP Lesion Visualizer.

PAD-UFES-20 [7]: a skin lesion dataset containing 2,298 close-up clinical images collected using different smartphone devices. It includes six types of skin lesions, data from 1,373 patients, and up to 22 clinical features per sample, covering both patient and lesion attributes, such as age, skin lesion location, and lesion diameter. The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Melanoma (MEL), and Nevus (NEV).

HIBA [27]: a skin lesion archive with clinical and dermoscopic images collected in Argentina, containing 1,616 images of 10 different types of skin lesions, including Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Melanoma (MEL), Nevus (NEV), Vascular Lesion (VASC), Lichenoid Keratosis (LK), Solar Lentigo (SL), and Dermatofibroma (DF).

HAM10000 [8]: also known as "Human Against Machine with 10,000 training images," this dataset comprises 10,015 multi-source dermoscopic images of skin lesions divided into seven classes and includes four clinical features, with two related to patient demographics and two describing lesion characteristics. The skin lesions are: Actinic Keratosis and Intraepithelial Carcinoma (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevi (NV), and Vascular Lesions (VASC).

PH2 [6]: a small dataset with 200 dermoscopic skin lesion images, including three classes: 80 common nevi, 80 atypical nevi, and 40 melanomas. The dataset contains 13 clinical lesion features, such as clinical and histological diagnosis, and the assessment of various dermoscopic criteria.

SLICE-3D [5], being the largest and most complete one, is considered as the reference dataset for pre-training the SLIMP model. All other datasets are considered as target datasets for performing skin classification using the pretrained model. Unless otherwise stated, evaluation is performed considering binary classification targets (benign/malignant) of the datasets that are better balanced.

For **PAD-UFES-20** [7], malignant classes include Basal Cell Carcinoma (BCC), Melanoma (MEL) and Squamous Cell Carcinoma (SCC), while benign classes include Actinic Keratosis (ACK), Nevus (NEV) and Seborrheic Keratosis (SEK). In **HAM10000** [8], Basal Cell Carcinoma (BCC) and Melanoma (MEL) are categorized as malignant, with benign classes comprising Actinic Keratosis (ACK), Nevus (NEV), Vascular Lesion (VASC), Dermatofibroma (DF), and Benign Keratosis-like Lesions (BKL). In **HIBA** [27], the malignant class includes Basal Cell Carcinoma (BCC), Melanoma (MEL) and Squamous Cell Carcinoma (SCC), while benign lesions encompass Actinic Keratosis (ACK), Dermatofibroma (DF), Lichenoid Keratosis (LK), Seborrheic Keratosis (SEK), Nevus (NEV), Vascular Lesion (VASC), and Solar Lentigo (SL). In the case of **PH2** [6] dataset, the malignant category consists only of melanomas, while common nevi and atypical nevi were grouped as benign. **SLICE-3D** [5], the largest dataset in this study, is inherently binary, with an extremely imbalanced distribution: 99.9% of lesions are benign, while only 0.1% are malignant.

C Extended ablation

We report additional ablations concerning the choice of image and tabular encoders, as well as the patient batch size. In the tables below, we highlight in light blue the reference configuration adopted in the experiments of the main text.

C.1 Image encoder

We consider the influence of the image encoder size on the downstream skin lesion classification task. Specifically, we consider the Tiny, Small & Base ViT variants [28, 37]. Table 7 shows the influence of the image encoder size on the performance metrics across four datasets: PAD-UFES-20, HIBA, HAM10000, and PH2. Interestingly, the influence of the image encoder size in the case of SLIMP is reduced, which can be attributed to the complementary information added by the metadata through the tabular encoder. Table 8 reports the number of parameters for the different image encoder sizes, with ViT-Base being approximately $4\times$ larger than ViT-Small and $15\times$ larger than ViT-Tiny.

		PAD-UFES-20				**HIBA**				**HAM10000**				**PH2 - w/out patient metada**			
		Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC
linprob	SLIMP w/ ViT-T	89.6	89.0	0.908	0.922	89.5	89.3	0.904	0.939	84.7	**81.7**	**0.665**	**0.910**	**95.0**	91.7	0.909	**1.000**
	SLIMP w/ ViT-S	**90.9**	**90.2**	**0.921**	**0.926**	**90.7**	**90.6**	**0.914**	**0.947**	**85.9**	78.5	0.650	0.901	**95.0**	**96.4**	**0.923**	0.988
	SLIMP w/ ViT-B	87.8	86.9	0.896	0.899	83.3	83.0	0.851	0.918	81.7	72.4	0.553	0.862	90.0	83.3	0.800	**1.000**
kNN	SLIMP w/ ViT-T	81.7	81.4	0.837	0.858	83.3	83.2	0.842	0.904	85.7	74.9	0.612	**0.904**	90.0	83.3	0.800	**1.000**
	SLIMP w/ ViT-S	**89.6**	**89.6**	**0.904**	**0.927**	**87.7**	**87.5**	**0.886**	**0.924**	**85.9**	**75.2**	**0.618**	0.888	**95.0**	**96.4**	**0.923**	0.988
	SLIMP w/ ViT-B	84.8	84.4	0.865	0.900	81.5	81.3	0.830	0.887	82.3	64.4	0.438	0.851	80.0	66.7	0.500	**1.000**

Table 7: Impact of image encoder size on the skin classification performance using SLIMP. Best results in **bold**.

The choice of N, the number of images and lesions selected per patient during training, also plays a role in performance differences. For ViT-Tiny and ViT-Small, $N = 100$ was chosen to balance computation and training efficiency, while for ViT-Base, $N = 50$ was used due to the model's significantly larger size and computational requirements. This may partially explain the performance drop observed in ViT-Base architectures, as the model has less diverse per-patient data for training.

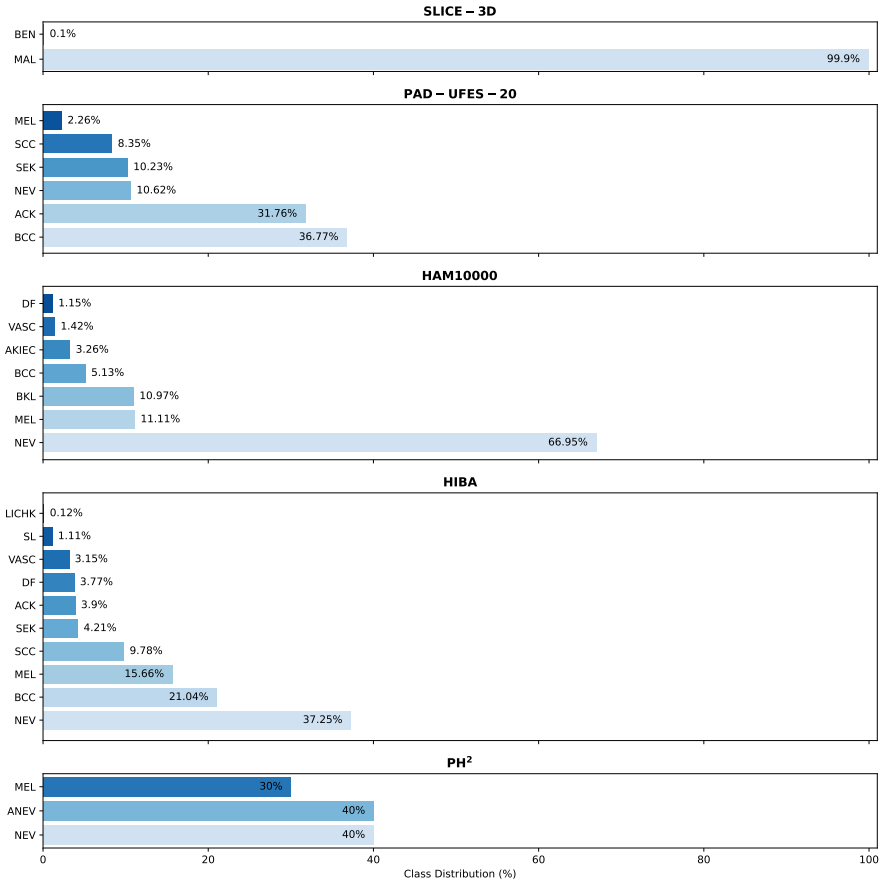

Figure 3: Representation of class distribution within each dataset considered.

	# of params (milions)			
		w/ TRACE		w/ FT-Transformer
	ViT-Tiny	ViT-Small	ViT-Base	ViT-Small
		SLIMP		
SLICE-3D	8.7	34.3	136	99.9
		SLIMP		
PAD-UFES-20	2.2	8.3	32.6	
HIBA	2.1	8.0	31.3	78.5
HAM10000	2.1	8.0	31.3	

Table 8: Number of parameters for the SLIMP and the SLIMP methods for different image and tabular encoders.

In summary, ViT-Small tends to strike the best balance between performance and model complexity, as seen across most datasets.

C.2 Tabular encoder

We compare the performance of SLIMP considering two tabular encoders: FT-Transformer [38] and TRACE [29]. Table 9 presents the corresponding performance across all datasets, using ViT-Small as the image encoder. TRACE, which is specialized for clinical data, consistently outperforms the generic FT-Transformer across all datasets and metrics considered, despite the fact that SLIMP with FT-Transformer has a significantly larger number of parameters, as shown in Table 8. In fact, despite being over four times bigger, FT-Transformer does not achieve the same level of performance. Moreover, in contrast to the adopted tabular encoder TRACE, FT-Transformer requires a significant

amount of hyper-parameter tuning to achieve optimal performance. These observations suggest that the task-specific design of TRACE offers a better balance of efficiency and performance when working with medical metadata, making it a more suitable choice for SLIMP.

		PAD-UFES-20				HIBA				HAM10000			
		Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC
linprob	SLIMP w/ FT-Transformer	89.6	89.1	0.908	**0.946**	84.6	84.0	0.871	0.910	80.2	50.0	0.000	0.655
	SLIMP w/ TRACE	**90.9**	**90.2**	**0.921**	0.926	**90.7**	**90.6**	**0.914**	**0.947**	**85.9**	**78.5**	**0.650**	**0.901**
kNN	SLIMP w/ FT-Transformer	87.4	87.2	0.886	**0.939**	82.7	82.6	0.837	0.882	77.7	52.4	0.159	0.745
	SLIMP w/ TRACE	**89.6**	**89.6**	**0.904**	0.927	**87.7**	**87.5**	**0.886**	**0.924**	**85.9**	**75.2**	**0.618**	**0.888**

Table 9: Comparison between the generic tabular encoder FT-Transformer and the tabular encoder for medical data TRACE. Best results in **bold**.

Table 10 compares the computational complexity, measured in GFLOPS, for SimCLR, SLIMP with FT-Transformer, and SLIMP with TRACE with different encoder sizes (ViT-Tiny, ViT-Small, ViT-Base). Naturally, computational costs scale with the size of the ViT encoder, highlighting the trade-off between model size and efficiency. In relation to metadata encoding, SimCLR which lacks metadata encoding is slightly more efficient with respect to the proposed multimodal SLIMP method, but SLIMP generally performs better, as has been shown in the results presented in the main text. On the other hand, the FT-Transformer tabular encoder introduces a significant overhead. The reference configuration featuring SLIMP with TRACE is a more balanced choice, offering improved performance with significantly less GFLOPS compared to the FT-Transformer. The number of GFLOPS for the supervised approaches SBCL, SBCL-C and TFormer are also reported in the table for comparison. Additionally, Table 11, reports the number of parameters and the relative training time between the SimCLR, SLIMP, SBCL and TFormer. Relative training times are normalized with respect to the SimCLR's training time on SLICE-3D.

	GFLOPS
SBCL(-C)	0.564
TFormer	4.509
SimCLR	1.258 | 4.608 | 17.582 (ViT-T | ViT-S | ViT-B)
SLIMP w/ FT-Transformer	1.694 | 6.298 | 24.233 (ViT-T | ViT-S | ViT-B)
SLIMP w/ TRACE	1.298 | 4.765 | 18.203 (ViT-T | ViT-S | ViT-B)

Table 10: Comparison of computational complexity in terms of GFLOPS between SBCL(-C), TFormer, SimCLR, SLIMP with FT-Transformer, and SLIMP with TRACE with different encoder sizes. ViT-T, ViT-S and ViT-B correspond to ViT-Tiny, ViT-Small and ViT-Base, respectively.

		SLICE-3D	PAD-UFES-20	HIBA	HAM10000	PH2
# params	SimCLR	5.5M				
	SLIMP	34.3M	8.3M	8.0M	8.0M	4.1M
	SBCL	0.5M	0.5M	0.5M	0.5M	0.5M
	TFormer		27.8M	27.8M	27.8M	27.8M
rel. time	SimCLR	1				
	SLIMP	0.3	0.04	0.03	0.1	0.002
	SBCL	0.2	0.06	0.05	0.01	0.002
	TFormer		0.01	0.01	0.04	0.002

Table 11: Model size comparison based on the total trainable parameters for every dataset (columns) and the relative training time, normalized to SimCLR's training time on SLICE-3D.

C.3 Patient batch size

We examine the impact of the patient batch size considered in the continual pre-training of the SLIMP on the PAD-UFES-20 dataset. Table 12 shows how the patient batch size affects performance on binary skin lesion classification. We observe that smaller batch sizes, as $B = 4$ and $B = 8$, yield slightly lower Balanced Accuracy (BAcc) and F1 scores, while larger batch sizes, lead to improved performance across all metrics but AUC. $B = 64$ achieves the highest BAcc of 89.5% and an F1 score of 0.913. Interestingly, further increasing the batch size (e.g., $B = 128$ or $B = 256$) does not

result in further performance gains and, in some cases, slightly decreases accuracy and F1 scores. This further highlights the importance of carefully choosing the patient batch size considered in the pre-training, as it can significantly impact performance. The choice of $B = 64$ strikes an effective balance, justifying its choice as the reference configuration.

	Acc	BAcc	F1	AUC
SLIMP w/ $B = 4$	90.0	86.4	0.886	0.907
SLIMP w/ $B = 8$	89.1	88.4	0.906	0.911
SLIMP w/ $B = 32$	88.7	88.4	0.898	**0.928**
SLIMP w/ $B = 64$	**90.9**	**90.2**	**0.921**	0.926
SLIMP w/ $B = 128$	89.6	89.1	0.908	0.918
SLIMP w/ $B = 256$	89.6	89.1	0.908	0.927

Table 12: Performance of the SLIMP method with different batch sizes (B) during the continual self-supervised learning stage on the PAD-UFES-20 dataset. Best results in **bold**.

D Textual data

We reproduce a concept-based interpretability (CBI) method [39], by adapting CLIP on the SLICE-3D dataset, considering a ViT-B/16 backbone architecture which offers optimal results. This methodology uses visual-language models for exploiting textual concepts for melanoma classification offering three different variants; (1) the *Baseline* approach, which directly applies CLIP, selecting the label that achieves the highest cosine similarity between the image and text embeddings, (2) the *CBM* approach, which introduces dermoscopic concepts and utilizes melanoma-specific coefficients to make predictions and (3) the *GPT-CBM* approach, which extends each dermoscopic concept introduced in CBM with multiple textual descriptions by querying it into ChatGPT.

In Table 13 we compare the performance of the above approaches, with our proposed SLIMP method, across three different target datasets, in a 'melanoma vs all' classification scenario. SLIMP is only adapted during linear probing while all pre-trained models on SLICE-3D dataset remain unchanged, highlighting the robustness of the learned representations. SLIMP consistently outperforms all other approaches without the need of task-specific pre-training.

	PAD-UFES-20				**HIBA**				**HAM10000**			
	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC	Acc	BAcc	F1	AUC
Baseline	23.9	51.3	0.044	0.422	68.5	54.8	0.261	0.502	72.0	58.6	0.247	0.595
CBM	78.7	69.6	0.109	0.778	48.2	61.3	0.333	0.659	54.1	58.8	0.238	0.565
GPT-CBM	35.7	57.3	0.051	0.599	48.8	61.7	0.336	0.638	55.5	57.6	0.231	0.581
SLIMP	**98.7**	**70.0**	**0.571**	**0.993**	**90.1**	**72.3**	**0.600**	**0.939**	**89.1**	**67.9**	**0.452**	**0.892**

Table 13: Comparison of SLIMP method with CBI variants across three target datasets. Results for the proposed SLIMP method are obtained using a linear probing setting. Best results in **bold**.

E Additional training details

Batch sampling strategy For both the initial and continual self-supervised pre-training stages, we construct each batch with B patients, including their respective patient-level tabular metadata. Additionally, for each patient, we sample N lesion images and their corresponding lesion-level tabular metadata. The number of lesions N varies per patient and is capped by an upper limit N_{max}. If a patient has more lesions, then a subset of $N = N_{max}$ lesions is randomly sampled in each epoch. In addition, a positive lesion sampling strategy is implemented, ensuring that, if a patient has malignant lesions, they are always included in the N lesions sampled during training. This ensures that the model encounters an adequate number of malignant lesions.

For the pseudo-modalities retrieval setup, where the images from the target dataset lack both lesion and patient metadata, we create two independent pools with tabular features derived from the metadata of the SLICE-3D reference dataset, by passing them through the pre-trained inner and outer tabular encoders. This step does not preserve any association between patients and their corresponding lesions. Consequently, the retrieval process of patient/lesion-level metadata is not constrained to

select features from the same patient across every modality, maximizing the flexibility of the proposed architecture.

Training details of supervised methods We pre-train SBCL [32] with a ResNet-32 architecture, for 1000 epochs on SLICE-3D dataset, followed by a dataset-specific continual pre-training (SBCL-C) for 100 epochs. Both pre-training setups use the SGD optimizer with a learning rate of 0.5 for the initial pre-training and $1e^{-2}$ for the continual pre-training. We evaluate each target dataset on the corresponding SBCL-C model, by applying linear classification for 150 epochs (following the SLIMP linear probing setting) with a learning rate of 0.1. During linear classification we select the Classifier-Balancing (CB) [40] train rule, which proved to outperform LDAM (Label-Distribution-Aware Margin Loss) [41].

Regarding TFormer [36], we utilize the variant designed to process two modalities, namely clinical images and tabular metadata, since the target datasets do not explicitly provide clinical and dermoscopic image pairs of the same lesion. During training, TFormer was fine-tuned on each target dataset, using Adam optimizer with a learning rate of $1e^{-4}$, and a weight decay of $1e^{-4}$. The learning rate was adjusted dynamically through the Cosine Annealing learning rate scheduler. The loss function used throughout the training process was Binary Cross-Entropy.

F Qualitative assessment

Figure 4 shows the t-SNE [42] embeddings of the three SLIMP variants presented in Table 5, on the PAD-UFES-20 dataset. We observe a better separation between benign and malignant lesions when metadata are considered during pre-training.

Figure 5 presents randomly selected lesions from each dataset with the corresponding attention maps extracted from the pre-trained image encoders of SLIMP, SimCLR (pre-trained on SLICE-3D under the same setting as SLIMP), DINOv2 and MAE, in this order. We note that SLIMP effectively localizes the majority of the lesions, regardless of differences in lesion shape, texture and color. This consistency in identifying relevant lesion regions indicates the robustness of the learned representations across diverse datasets that exhibit a high variation in visual appearance. It also showcases the ability of the model to focus on important visual features, supporting the improved downstream classification performance, and suggesting that the method can enhance the interpretability and reliability of the results.

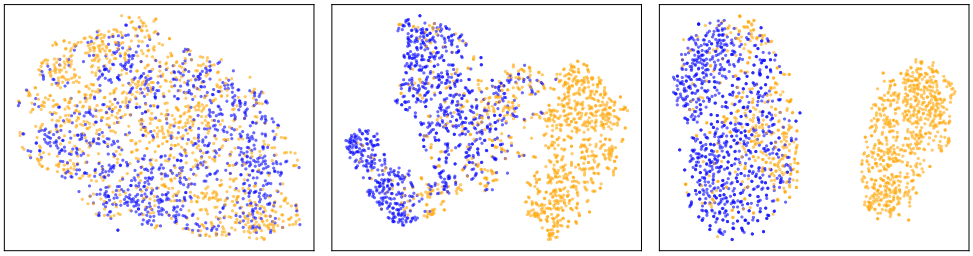

Figure 4: t-SNE visualization of SLIMP features for benign and malignant lesions in the PAD-UFES-20 dataset. **Left:** Pre-training using image encoder alone; **Middle:** Pre-training using image and lesion metadata; **Right:** Pre-training using images with lesion and patient-level metadata.

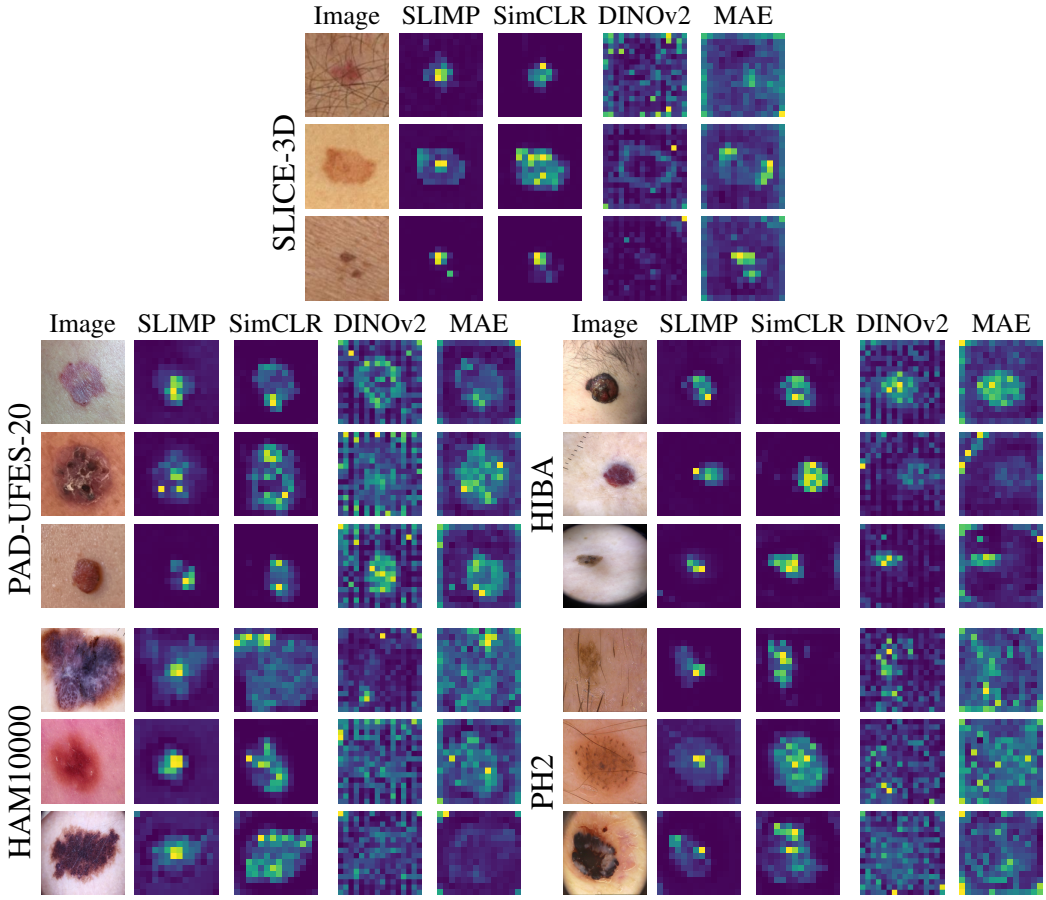

Figure 5: Attention maps of SLIMP as obtained by the image encoder's last block, in comparison with the attention maps of SimCLR, DINOv2 and MAE pre-trained image encoders for the SLICE-3D source dataset (top), and across four target datasets; PAD-UFES-20 (middle-left), HIBA (middle-right), HAM10000 (bottom-left) and PH2 (bottom-right).